Data Mining for Material Flow Analysis: Application in the Territorial Breakdown of French Regions

Brinduşa Smaranda

Master DMKM Report

Data Mining for Material Flow Analysis: Application in the Territorial Breakdown of French Regions

Brindusa SMARANDA

defended the 07/09/2013

Supervision : Emmanuel Prados, Jean-Yves Courtonne, Peter Sturm - STEEP Team (INRIA) and Tomàs Aluja - Universitat Politècnica de Catalunya

Location : STEEP Team, INRIA Rhône-Alpes

Abstract:

*One of the major issues for assessment of the long-term sustainability of urban areas is related to the concept of "imported sustainability". In order to produce such an assessment for a given territory, one must first identify and quantify the types of materials used, and the impacts associated to these uses. Material Flow Analysis (MFA) is directly related to how the material circulates and how it is transformed within a territory. In most cases this analysis is performed at national and regional levels, where the statistical data is available. The challenge is to establish such an analysis at smaller scales, e.g. in the case of France, at the department or city level. Currently, few studies are done at these scales and when they exist, they are based on the extrapolation of data at the country or the region levels. In this thesis, the possibility of applying data analysis at the regional level by generating a mathematical model that can fit well the data at regional scale and estimate well the departmental one is explored. The downscaling procedure relies on the assumption that the obtained model at level **'n'** (for example region) will be also true at level **'n+1'** (for example department), such that it could properly estimate the unknown data based on a set of chosen drivers. The tests show that it is very important to choose the proper drivers and the class of model.*

Résumé :

*Une des problématiques les plus importantes dans l'évaluation de la durabilité des zones urbaines est liée au concept de "durabilité importée". Pour produire une telle évaluation d'un territoire donné, il est dans un premier temps nécessaire d'identifier et de quantifier les flux de matière mobilisés par le territoire ainsi que les pressions environnementales associées à ces flux. L'Analyse de Flux de Matière (AFM) est directement liée à la façon dont les matériaux circulent et sont transformés par le système socio-économique. Dans la plupart des cas, cette analyse est réalisée à l' échelle nationale ou régionale, où les statistiques sont disponibles. Le défi consiste à établir de telles analyses à des échelles plus fines, par exemple, dans le cas de la France, à l'échelle des départements ou encore des villes. Actuellement, peu d'études existent à ces échelles et lorsque c'est le cas, elles comprennent l'extrapolation des données sur le pays ou le niveau de la région. Dans cette thèse, la possibilité d'appliquer une analyse au niveau régional par la génération d'un modèle mathématique qui peut s'adapter à bien des données à l'échelle régionale tant que estimer bien les départementale données est explorée. La méthodologie de descente d'échelle repose sur l'hypothèse que le modèle obtenu au niveau **'n'** (par exemple région), sera toujours vérifié au niveau **'n+1'** (par exemple département), de telle façon qu'il soit possible d'estimer une donnée inconnue sur la base d'un jeu de variables explicatives connues. Les tests realisés soulignent l'importance d'une selection pertinente des variables explicatives ainsi que de la classe du modéle.*

Contents

1 Hosting institution

1.1 INRIA

INRIA or "**I**nstitut **N**ational de **R**echerche en **I**nformatique et **A**utomatique" is a public science and technology institution established in 1967 in France. INRIA is the only French public research body fully dedicated to computational sciences [6]. INRIA promotes constant exchanges between international specialists in research and development.

The national institution has partnerships with many academic institutions in Europe, United States, Asia, Latin America, Africa, and the Middle East. Moreover it has joint laboratories with University of Illinois at Urbana-Champaign and with Academy of Sciences in China. It has a common laboratory with Microsoft Research and with Alcatel-Lucent Bell Labs. Also, the French office of World Wide Web Consortium (W.3.C.) is hosted by INRIA. Furthermore a center for research and innovation has been created in Chile.

INRIA is a founding member of ERCIM (European Research Consortium for Informatics and Mathematics), which brings together 20 European research institutes [5]; the research is organized in teams of 10 – 30 people.

1.2 STEEP Team

"**S**ustainability **T**ransition, **E**nvironment, **E**conomy and local **P**olicy" or STEEP [4], lead by Emmanuel Prados, is an interdisciplinary research team of INRIA Rhône-Alpes and LJK (Laboratoire Jean Kuntzmann).

This laboratory collaborates with Université Joseph Fourier from Grenoble and with CNRS (Centre National de la Recherche Scientifique). STEEP is devoted to systemic modeling and simulation of the interactions between the environmental, economic and social factors in the context of transition to sustainability at local (sub-national) scales.

Its objective is to set up some mathematical and computational concepts to develop decision-making tools.

Although STEEP is newly created team, in 2012 it was present at the conference "Flow Modeling for Urban Development" organized under the aegis of French research network "Groupement d'Intérêt Scientifique" on "Urban Modeling".

The team ongoing projects are as follows. CITiES aims to define models for the design and evaluation of land use and transportation policies (LUTI). The goal of ESNET project is to assess alternative futures of ecosystem services networks for the urban area of Grenoble city. TRACER is a scheme program, which quantifies elements of urban dynamics necessary to implement policies that are coherent with sustainable urban objectives.

This thesis presents the work I performed within the STEEP related to data mining for material flow analysis as a support of territorial metabolism understanding.

2 Acknowledgement

I would like to thank the entire STEEP team for their cooperation and collaboration. I am grateful for the unconditional support and help given by the researchers Emmanuel Prados and Peter Sturm together with the PhD student Jean-Yves Courtonne, who reviewed and encouraged my ideas.

I would like also to thank Prof. Tomàs Aluja for his prompt answers and for his comments in shaping the thesis. Besides being my teacher and my master thesis supervisor, Prof. Aluja is a very good coordinator and together with all the teachers from Universitat Politècnica de Catalunya, Barcelona (Spain), provided us a great academic experience. Thank you for the great assistance and for your true interest in transmitting your knowledge to us!

I must address a sincere thank you to many of my DMKM colleagues, who supported me in different ways and provided with a wonderful environment for the past two years.

Finally I cannot thank enough my family and I would like to dedicate this work to my parents, who have been there all along to take care of me and to morally support me.

3 Introduction

3.1 Territorial Metabolism

Urban areas are characterized by a concentration of economic activities, a large population, and large material stock densities, inducing high levels of energy and material flows [22]. The main types of material within a territory are classified into four categories: biomass, metals, minerals and fossil fuels. These material flows represent potential ecosystem impacts on different scales, ranging from local to regional up to global [9].

To advance long-term sustainability of urban areas it is vital to understand the region's metabolism, and this requires a detailed knowledge of the main material flows. According to [24] this management should mainly include the quantification of material flows.

Studying the metabolism of a territory consists in analyzing the interrelations between the economy and the environment, in which economy works like an environmental subsystem, dependent on the ongoing throughput of materials and energy [17]. Therefore, raw materials, water, air, gas, cereals, etc. are extracted from the natural system, constituting inputs to the economic system, and are partially transformed into products, residues, and other material and energy flows that may cause environmental damage.

For example in Figure 3.1 it is presented the Sankey diagram, which depicts the flows of different types of cereals from the extraction stage until they become final products delivered to customers. This representation is helpful for the evaluation of material interaction with the environment, and also to understand how much a territory depends on the others.

In order to build such a diagram for sub-national territories, one must appreciate the material consumption of a system for a certain year. In [26], Niza et al. speak about MFA (Material Flow Analysis), which is considered as being a tool that can simultaneously disaggregate the data and characterize the dynamics of the metabolism of an area.

3.2 Material Flow Analysis

Material flow analysis (MFA) facilitates the assessment of a system's material consumption for a certain period, generally one year, but also allows the evaluation of trends in material consumption of the economic system through the development of time series.

In this context, MFA can support decision makers in coming to understand the metabolism of a region.

More specifically, MFA examines the materials flowing into a given system (private household, company, region, city, etc.), the stocks and flows within this system, and the resulting outputs from the system to other systems (export, wastes, etc.).

The MFA scheme (Figure 3.2) can be seen as a system, which receives some inputs, namely imports and local extractions, stores the needed material for internal use and outputs the emissions/wastes together

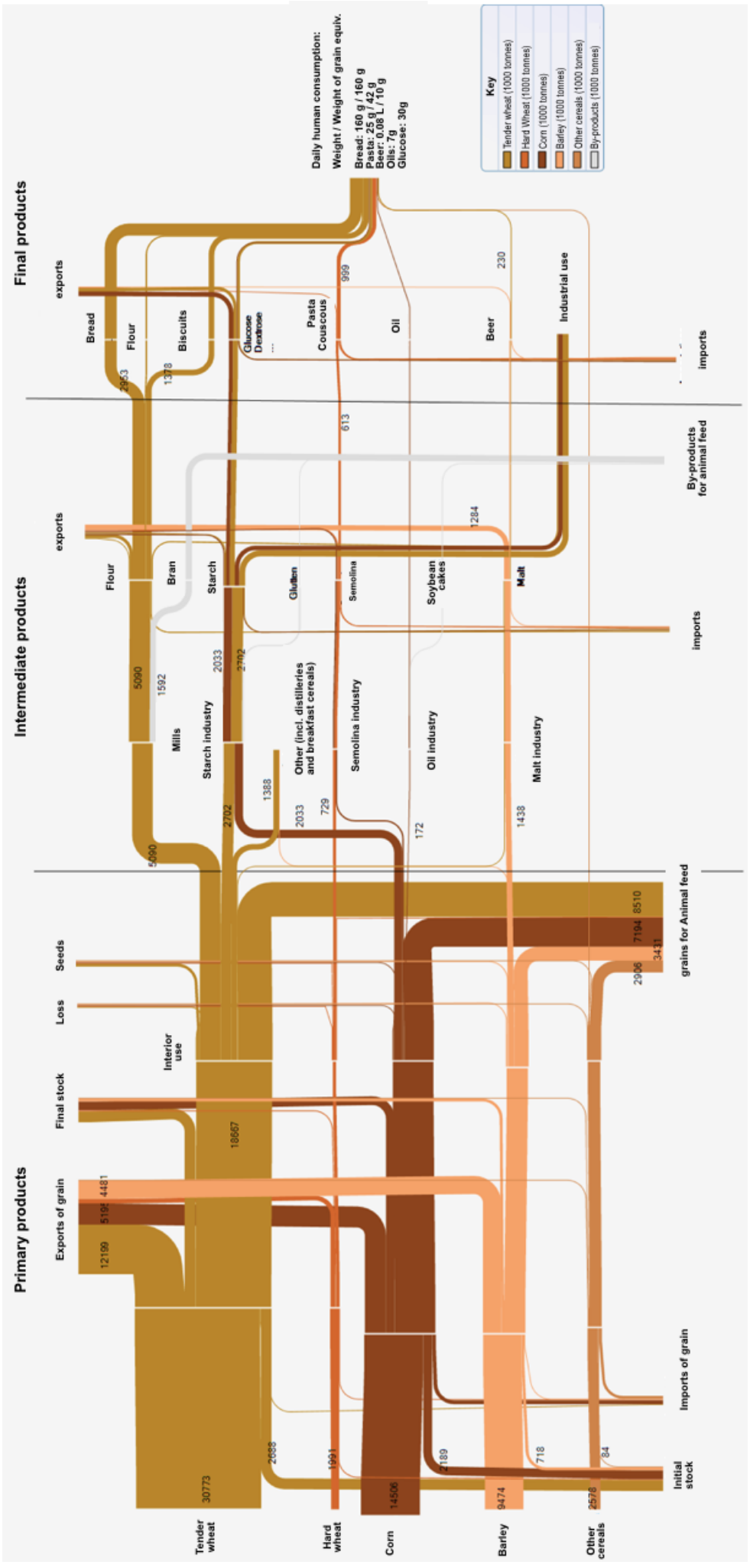


Figure 3.1: Cereals Flow – Sankey Diagram

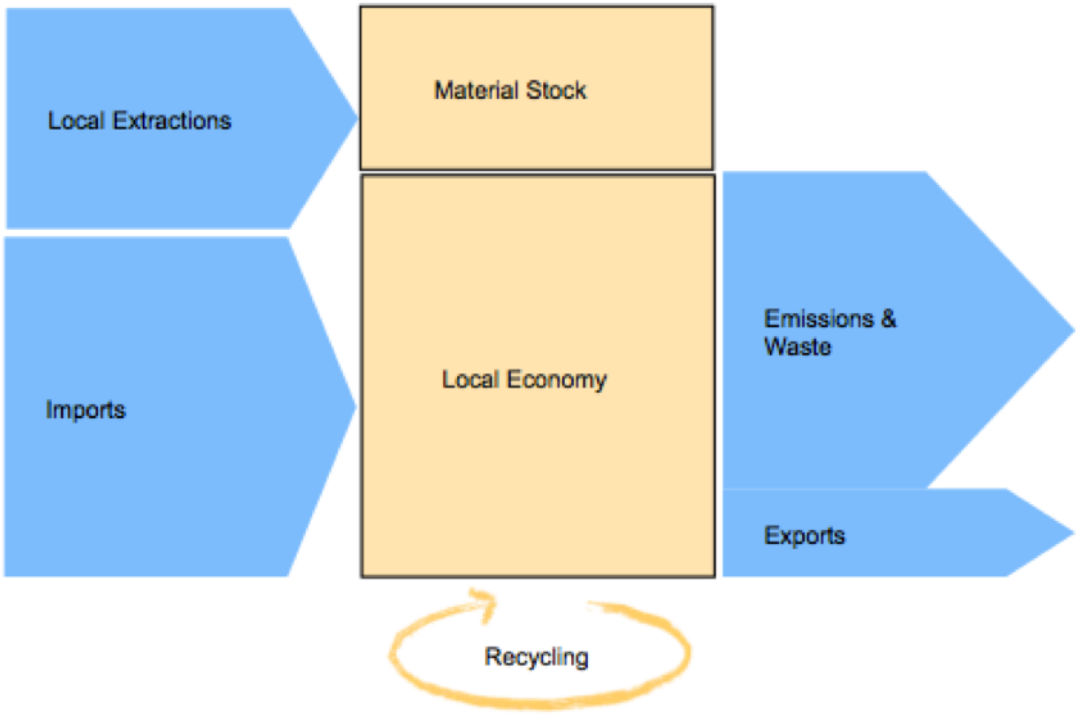


Figure 3.2: Material Flow Analysis Scheme

with the export of different products. What is not known is the local economy and the goal is to understand its process inside the system based on the given inputs and outputs. Therefore, following the example of the cereals, one would be interested in describing the local economy as a diagram, which combines the total production of the cereals with its transformations and with the finite products that can be obtained such as bread, biscuits, etc. In other words, a Sankey diagram [31] will be the perfect tool for this analysis and it should be accessible for all levels of a territory (national, regional and local).

MFA is useful for examining the relationship between a region or city and its surrounding hinterland [28]. Nevertheless, as Binder points out in her review of regional MFA [12], there is no methodological framework for this type of study, nor are there suitable data. If I look at urban MFA this is true as well, and the quantity of available data is even less.

3.3 The Problem

A large number of MFA studies have been done at a national level, but studies focusing on the regional or local level have still been very limited, and a standardized method equivalent to that presented by EUROSTAT [30] for the national level does not yet exist. Although the available studies show the importance of material flows for the regional and urban metabolism, they also present a large spectrum of approaches that can be defined through the MFA approach [22].

The main obstacle to attempt to calculate the material flow of a city or region is the lack of statistical data.

This calls for different approaches to assessing urban material flows. Studies tend to either focus

on choosing and analyzing only the most important products and materials (Bünz Valley [14], Greater London [11], Paris [10], Geneva [19]) or focus on tracing a specific substance, such as lead, copper, or phosphorus (Bünz Valley).

Several methodological solutions are available in the literature and include extrapolation of data from the country or the region and estimations based on sales, number of inhabitants or workers.

The purpose of this study is to establish a more principled procedure for solving the lack of data in a specific territory. Hence, the basic idea is to apply a downscaling technique on the available data. A possible analysis in case of France is to estimate these data for the department scale, knowing the regional one. In other words, the problem can be formulated as follows: using the statistical data of the regional level, for example the production of cereals in kilotons, I would like to estimate the departmental cereal's production. In this sense, I will use the other known statistical data such as surface of cultivated land, the number of workers in the cereal's industry, the population, etc. by calling them drivers and I will check if these drivers can explain the production of cereals at region level and if the relationship between them can be kept also at department level.

Following this approach, I could automatize the process for obtaining reliable data for different categories of materials at department level. In the same time based on the accuracy of each model, it can be understood if the adopted policy is good or not. On the other hand the verification process is very easy to be conducted. Therefore, if the data is accessible at department level, then I can compare it with the estimated data and check how well the model performs.

This work is organized as follows: in section 4, I present the state of the art in the field of material flow analysis, by underlining several data quantification concepts. The used methods are based on extrapolation or on very simple mathematical computations, without considering the correlation between different materials and possible drivers. The conclusion is that as long as within a territory the production of materials depends on different drivers, it is interesting to study the correlation between them and finally to find a linear model that will best estimate the data.

Section 5 describes the implemented methodology and offers a mathematical demonstration of the solution for the stated problem. The proposed approach integrates linear constrained optimization with linear models.

Section 6, practical part, presents a case study in the territorial breakdown of French regions. Challenges that are managed in this part are linked to model and drivers selection, but also to overfitting and outlier handling. Because the number of drivers is larger than that of observations, it is difficult to fit a model, and the problem of overfitting must be controlled. As a first step, I decided to manually choose the drivers in order to not allow performing a wrong analysis. An AIC-based criterion [8] within a RANSAC (RANdom SAmple Consensus) [21] procedure it is used to select the model taking in consideration that data have outliers.

The conclusions are given in section 7.

In section 8, I present the outgoing and the future work of the project. Another technique based on Bayesian Networks will be studied and applied on the presented problem.

4 State of the Art

Pioneering studies in the MFA field highlight the importance of territorial metabolism and provide some data on the urban material balance [33], [29], [13]. These studies, however, are presented more often at national scales and do not directly apply at local level. More recently, new methodological developments and case studies provide encouraging results and pave the way for further studies. MFA has now been shown to be relevant not only in describing socio-natural interactions but also in supporting public policies and action (see for instance for Vienna: [18] for Stockholm:[15]; for Geneva: [20]; for Hamburg, Vienna, and Leipzig: [23]).

Nevertheless, studies focusing on the regional or local level have still been very limited, when compared to the large number of MFA studies performed on the national level. Moreover, standardized methods equivalent to the one presented for the national level by EUROSTAT [30] does not yet exist. Although the available studies show the importance of material flows to regional and urban metabolism, they also present a large spectrum of approaches that can be defined through the MFA approach [22].

The lack of available statistical data at the municipal and regional levels currently calls for different approaches to assessing urban/local material flows. Some of the studied methodologies focusing on urban areas are summarized in Table 4.1.

Studies tend to either focus on choosing and analyzing only the most important products and materials based on resource flows and ecological footprint (Greater London, Geneva, Paris) or focus on tracing a specific substance, such as lead, copper, or phosphorus, by gathering data from similar regions as the ones that are studied (Bünz Valley).

Another solution for data quantification has been presented in the case study of Lisbon. In this study, S. Niza et al. (2009) quantify the Lisbon's material balance for 2004 by using a system of matrixes composed of different calculated ratios. The idea is to build different matrices based on the available data at urban, regional and national scales.

Other authors managed to handle the lack of available statistical data at low-level scales, by using a technique that breaks down the data from upper levels (national, regional). This method has been applied on the case study of Czech Republic in 2009. In this study, the breakdown method is based on territorial distribution and on the total amount of existing materials. In order to obtain the data at the level of region some basic mathematical computations have been applied on the data from upper levels. A good example is the following one: if a certain region represents 20% of the surface where a specific material is produced, then the amount of material production for that region is assumed to be 20% of the entire production at the national level.

All these approaches lead to studies that generally do not explain the correlation between material and drivers within a region or city and also they have not been tested sufficiently.

Although the correlations between the materials and drivers are not observed in too much detail, a study that analyzes them has been done within the STEEP team and it is presented in the master thesis of Jean-Yves Courtonne (2011) [16]. In his thesis, Jean-Yves presents the correlation for different

The related study	The methodology used
Industrial metabolism at the regional and local level: A case-study on a Swiss region, 1994 [14]	Gathering data from similar regions as Bünz Valley and adapted them to identify specific substance
City limits: A resource flow and ecological footprint analysis of Greater London, 2001 [11]	Resources flow and ecological footprint assessment
Ecologie industrielle à Genève. [Industrial ecology of Geneva], 2005 [19]	Resources flow assessment
Methodological Advances in Urban Material Flow Accounting Based on the Lisbon Case Study, 2009 [27]	Data quantification by making use of a system of matrixes built up with the available data at the urban, regional and national scales: materials matrix, throughput matrix, waste treatment matrix and activity sectors matrix
Analysis of regional material flows: The case of the Czech Republic, 2009 [25]	Data breakdown based on territory distribution and on the total amount of existing material.
Urban Metabolism of Paris and Its Region, 2009 [10]	Identifying the principal material flows and using the Eurostat method updated to local or regional levels
Étude des flux de matiéres et d'énergie: région Rhône-Alpes, département de l'Isère, bassin d'emploi de Grenoble. [Material and energy flow study : Rhône-Alpes region, Isère department, Grenoble and its industrial area], 2011 [16]	Data correlation applied for different materials with possible drivers (eg: correlation between cereals production and particular types of land)

Table 4.1: Methodologies used to obtain MFA data

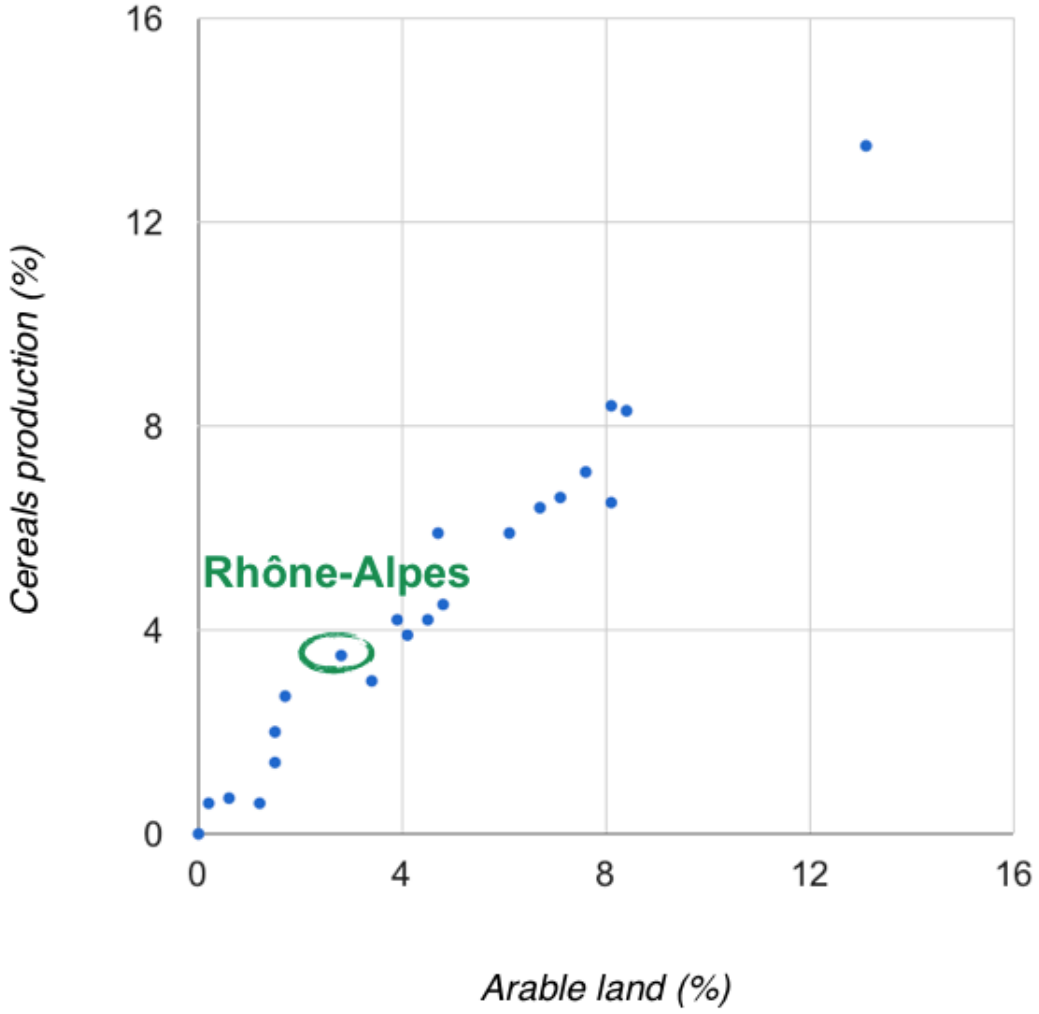
biomass extractions and possible drivers. For example, he studies the following relationships between: agricultural production and the added value in agriculture; agricultural production and its employment rate; agricultural production and the cultivated land. In his study, the used data sets are: the Corine Land Cover data [3]; the census data, which was performed by Agreste(2000) [1] for agricultural and economic sector (the added value of the agricultural sector, the number of farmers) (for more details please see the sub-section 5.2).

In Jean-Yves study, the observed correlation is of type 1 to 1, which means the correlation is between only one driver and a selected material.

Figure 4.1 presents the relationship between the production of cereals and arable land for region level, explained by a high correlation coefficient of 0.98. If the correlation would be perfect, then all the points would lie on one line. Each, cereal's production and arable land represents a certain percentage from the total national production and land surfaces.

In conclusion, the main methods for data quantification in studied literature refers to arbitrary ratios between different materials and drivers and to correlation studies of type 1 to 1 . Therefore, the idea is to study multiple correlations between materials and different drivers, under the assumption that they are maintained also at disaggregate levels. The latest idea refers to downscaling hypothesis, which states

Figure 4.1: The cereal production (%) explained as a function of the arable land (%) for French regions



that the explanatory model of materials remains the same when the scale changes. Based on the accuracy of the model, it can be deduced if the adopted policy is good or not. Also, the model shows how much the other variables explain the materials. Knowing the model and in some cases the data for department, I can test if the estimated data is the same as the real one.

My objective is to automatize the correlation analysis with many drivers, overcoming the problem of overfitting and outliers.

5 Methodology

In this section, I present the territorial breakdown methodology of French regions aiming to (1) identify the available data, (2) select the most appropriate drivers, (3) compute the best model, by overcoming the problem of outliers and overfitting.

5.1 The Context

The goal is to analyze an approach, which estimates the necessary data for departments and their subdivisions of Metropolitan France.

France is divided into 27 administrative regions, out of which 22 are in Metropolitan France (Figure 5.1), and five are overseas.

Figure 5.1: Regions of Metropolitan France

In the administrative division of France, the department (French: département) is one of the three levels of government below the national level ("territorial collectivities"), between the region and the commune. There are 96 departments in Metropolitan France, which are numbered from 1 to 95 and 5 overseas departments (Figure 5.2), which also are classified as regions. Departments are further subdivided into 342 arrondissements, themselves divided into cantons.

5.2 Mobilized Data

Most of the available data is at national and regional scales. However for some specific materials, data at department level is also accessible, such that the tests are focusing on this scale. Because different sources are involved, a preprocessing phase is necessary:

a) detect and extract the data at all three levels: national, regional, department;

Figure 5.2: Departments of France

b) organize them in **R** software environment format;

c) prepare them for modeling stage.

5.2.1 Sources

The Table 5.1 presents the sources that are used in the territorial breakdown of French regions analysis. In the case of materials, the data is classified based on the type of extraction and type of material. For the drivers, the sources are divided based on the type of data and its origin.

Data	Type	Source	Years	Lowest Geographical Level	Units
Materials	Agriculture and Sylviculture	Agreste [1]	1989 – 2010	Department	Physical Units (e.g.: Kilotons, m^3)
	Fisheries and aquaculture	France AgriMer [2]	2007, 2008	Region	Physical Units (e.g.: Kilotons, m^3)
Drivers	Population	INSEE [7]	1999, 2007, 2008	City	NA
	Employment by sector	INSEE [7]	2007	City	NA
	Biophysical Land Occupation	CORINE Land Cover [3]	1990, 2000, 2007	City	Hectares

Table 5.1: Mobilized Sources

In Table 5.1, "Years" column refers to the available years; the "Lowest Geographical Level" column represents the smallest scale for which data is feasible and the "Units" column expresses the units of measurement for each data. In the case of population and employment, the units represent the number of people for every territory or employment sector. Concerning the giving analysis, I take in consideration only the data from the year 2007.

The Corine Land Cover (CLC) [3] databases and several of its programmes have been taken over by the European Environment Agency (EEA) and it is produced in the framework of the European COoRdination of INformation on the Environment (CORINE). It represents an inventory of land cover for 44 classes (Table A.1 from Appendix A), and it is presented as a cartographic product, at a scale of

1 : 100000. This database is available for most areas of Europe. CLC is derived from visual interpretation of satellite images, with additional supporting data. The surface of the smallest unit mapped is 25 hectares. The land use classification was developed based on specific objectives: to map the entire territory of the European Union, to give the environment status and to not contain ambiguous positions. Therefore it is focused on the biophysical land and not on its use; it emphasizes the nature of objects (forests, crops, water bodies, rocky outcrops, etc.) rather than their socio-economic function (agriculture, housing, etc.).

The agricultural census conducted by Agreste (the statistical office of the Ministry of Agriculture and Fisheries) is performed every 10 years. The version available at this time is for 2010. The advantage of this database is the level of accuracy both geographically and in terms of cultures. Unlike the CLC data, the distinction is made between different types of crops (cereals, oilseeds, vegetables and potatoes, etc.).

France's National Institute of Statistics and Economic Studies (Institut National de la Statistique et des Études Économiques: INSEE) collects the data needed to compile quantitative results. It undertakes censuses and surveys, manages databases, and also draws on administrative sources. Each year, INSEE estimates population of regions and departments. INSEE's program for employment estimation, counts the number of people in employment residing in France for different employment sectors (cereals culture industry, mineral industry, etc.).

5.2.2 Data Normalization

The retrieved data have different units of measures. This requires data to be prepared before models can be built. In this context, preparing the data means transforming them prior to the analysis. One of the most basic transformations is normalization.

Basically, normalizing means transforming so as to render normal. When data are seen as vectors, normalizing means transforming the vector so that it has unit norm. The variables (columns) are normalized to the same 'dynamic range', with no units (they become a-dimensional values).

In the case of territorial breakdown of French regions, the variables are distributed with distinct values. Then, the idea is to express them as percentages of the total available data for a specific territory. For example, the production of cereals for region Rhône-Alpes represents 3.76% from the total amount of cereal's production at national scale, which is equivalent to 100%. In the computations, I used the interval $[0, 1]$ as evaluation scale of the data percentages. Therefore the cereal's production is 0.0376 of the total amount equal to 1.

5.3 Used Notations

In order to understand better the variables that I used in the methodology, I present the meaning of each of them:

- Y is the response variable (e.g.: the production of cereals);

- X^j the drivers (e.g.: the cultivated surface, the added value, etc.) and each of them is defined as X^1, X^2...;
- The prefix l identifies the data from a given level (e.g: r - region level; d - department level), where $Y_{l_1}, Y_{l_2}, \cdots, Y_{l_i}$ represent the response variables for i subdivisions of that level and $X^1_{l_1}, X^1_{l_2}, \cdots, X^1_{l_i}$ is the set of first driver for each subdivision. For example for the Rhône-Alpes region, Y_{l_1} represents the response variable for Isère department, Y_{l_2} is the response variable for Drôme department, and so on. In the same sense $X^1_{l_1}$ is the cultivated surface for Isère, $X^2_{l_1}$ is the added value for Isère;
- The estimated variables are denoted as $Y^*_{l_i}$.

5.4 Finding an Explanatory Model at the Regional Level

Let's consider the general linear function $Y^*_{r_i} = \lambda_1 X^1_{r_i} + \lambda_2 X^2_{r_i} + \cdots + \lambda_m X^m_{r_i}$ where m is the number of the drivers and r stands for region level. Therefore, the goal is to find the best estimation of coefficients λ_j, which can minimize the difference between estimated and true outcomes $Y^*_{r_i} - Y_{r_i}$ such that the preset condition $\sum_{i=1}^{n} Y^*_{r_i} = Y_{FR}$ is fulfilled, where Y_{FR} represents the total amount of a material at French national level and n is the number of local subdivisions (e.g.: n is the number of regions within a territory or it can denote the number of departments within a region, etc.). Because $Y_r = 1$ and $\sum_i X^j_{r_i} = 1$ (cf. sub-section 5.2.2), then the constraint is transformed into $\sum_j \lambda_j = 1$ (please see the proof in Appendix A).

Taking into consideration that the regional data is known, then the equation system of linear functions for the regional level is:

$$\begin{cases} Y_{r_1} = \lambda_1 X^1_{r_1} + \lambda_2 X^2_{r_1} + \cdots + \lambda_m X^m_{r_1} \\ Y_{r_2} = \lambda_1 X^1_{r_2} + \lambda_2 X^2_{r_2} + \cdots + \lambda_m X^m_{r_2} \\ \vdots \\ Y_{r_n} = \lambda_1 X^1_{r_n} + \lambda_2 X^2_{r_n} + \cdots + \lambda_m X^m_{r_n} \end{cases} \tag{1}$$

For this system the only unknown variables are lambdas, λ_j. This system has the number of equations larger than the number of unknown variables ($n > m$). Therefore there is no exact solution; so the method ordinary least squares is used to find the best approximation for the coefficients. In the end for the regional scale, the problem of finding the best approximated values for lambdas means to solve the following optimization problem:

$$\begin{cases} \min & \sum_{i=1}^{n} (Y_{r_i} - \lambda_1 X^1_{r_i} - \lambda_2 X^2_{r_i} - \cdots - \lambda_m X^m_{r_i})^2 \\ \text{with} & \sum_{j=1}^{m} \lambda_j = 1 \end{cases} \tag{2}$$

The Lagrangien function of the optimization problem is:

$[\sum_{i=1}^{n}(Y_{r_i} - \lambda_1 X_{r_i}^1 - \lambda_2 X_{r_i}^2 - \cdots - \lambda_m X_{r_i}^m)^2 + 2\mu(\sum_{j=1}^{m}\lambda_j - 1)]$

The system of partial derivatives becomes:

$$\left\{\begin{array}{lll} \frac{1}{2}\ \frac{\partial}{\partial\lambda_1}\ [\sum_{i=1}^{n}(Y_{r_i} - \lambda_1 X_{r_i}^1 - \lambda_2 X_{r_i}^2 - \cdots - \lambda_m X_{r_i}^m)^2 + 2\mu(\sum_{j=1}^{m}\lambda_j - 1)] & = & 0 \\ \frac{1}{2}\ \frac{\partial}{\partial\lambda_2}\ [\sum_{i=1}^{n}(Y_{r_i} - \lambda_1 X_{r_i}^1 - \lambda_2 X_{r_i}^2 - \cdots - \lambda_m X_{r_i}^m)^2 + 2\mu(\sum_{j=1}^{m}\lambda_j - 1)] & = & 0 \\ \vdots & & \\ \frac{1}{2}\ \frac{\partial}{\partial\lambda_n}\ [\sum_{i=1}^{n}(Y_{r_i} - \lambda_1 X_{r_i}^1 - \lambda_2 X_{r_i}^2 - \cdots - \lambda_m X_{r_i}^m)^2 + 2\mu(\sum_{j=1}^{m}\lambda_j - 1)] & = & 0 \\ \frac{1}{2}\ \frac{\partial}{\partial\mu}\ [\sum_{i=1}^{m}(Y_{r_i} - \lambda_1 X_{r_i}^1 - \lambda_2 X_{r_i}^2 - \cdots - \lambda_m X_{r_i}^m)^2 + 2\mu(\sum_{j=1}^{m}\lambda_j - 1)] & = & 0 \end{array}\right. \tag{3}$$

In order to solve this optimization problem, the following matrixes are used:

$$\mathbf{Y} = \begin{bmatrix} Y_{r_1} \\ Y_{r_2} \\ \vdots \\ Y_{r_n} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} X_{r_1}^1 & \cdots & X_{r_1}^m \\ X_{r_2}^1 & \cdots & X_{r_2}^m \\ \vdots & & \\ X_{r_n}^1 & \cdots & X_{r_n}^m \end{bmatrix} \text{ and } \lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix}$$

The resulting system is:

$$\left\{\begin{array}{ll} \min & (\mathbf{Y} - \mathbf{X}\lambda)^t(\mathbf{Y} - \mathbf{X}\lambda) \\ \text{with} & \mathbf{I^t}\lambda - 1 = 0 \end{array}\right. \tag{4}$$

Or:

$$\left\{\begin{array}{lll} \frac{1}{2}\ \frac{\partial}{\partial\lambda_i}\ (\mathbf{Y} - \mathbf{X}\lambda)^t(\mathbf{Y} - \mathbf{X}\lambda) + \mu(\mathbf{I^t}\lambda - 1) & = & 0 \quad \forall i = 1, \cdots, n \\ \frac{1}{2}\ \frac{\partial}{\partial\mu}\ \mathbf{I^t}\lambda - 1 & = & 0 \end{array}\right. \tag{5}$$

$$\left\{\begin{array}{lll} -\mathbf{X^t}(\mathbf{Y} - \mathbf{X}\lambda) + \mu I & = & 0 \\ \mathbf{I^t}\lambda - 1 & = & 0 \end{array}\right. \tag{6}$$

$$\left\{\begin{array}{lll} \mathbf{X^t}\mathbf{X}\lambda + \mu\mathbf{I} & = & \mathbf{X^t}\mathbf{Y} \\ \mathbf{I^t}\lambda & = & 1 \end{array}\right. \tag{7}$$

Using the matrices formulations the system becomes:

$$\begin{bmatrix} \mathbf{X^t}\mathbf{X} & \mathbf{I} \\ \mathbf{I^t} & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} \mathbf{X^t}\mathbf{Y} \\ 1 \end{bmatrix} \tag{8}$$

The solution of the system can be obtained by solving:

$$\begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} \mathbf{X^t X} & \mathbf{I} \\ \mathbf{I^t} & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X^t Y} \\ 1 \end{bmatrix} \tag{9}$$

Where μ represents the approximation error for coefficients.

The matrices $\mathbf{X^t X}$ and $\mathbf{X^t Y}$ represent the correlation matrices between the drivers and themselves respectively between the drivers and response variables. They can be calculated very easily, based on the given data.

5.4.1 Model and Driver Selection

The major challenges here are to select those drivers, which can describe very well each material and to find the best model that can fit the data.

In the case of agriculture and sylviculture, I select the following types of materials:

- cereals;
- fruits;
- potatoes;
- vegetables;
- grapes.

The drivers that I would like to include in the analysis are:

- population;
- employment for each sector;
- different types of land.

For France, the employment database differentiates between 709 sectors of activity and 44 types of CLC surfaces. Therefore the total number of drivers for this analysis is 753. The number of observations for the region is equal to 22 and for departments is 96, when the data is available. Taking in consideration the number of drivers is larger than that of observations, by fitting a model with all the drivers; I risk the statistical model to describe random error or noise instead of the underlying relationship. In order to control the model overfitting, I analyzed the correlation between the materials and the drivers. The results prove that I cannot totally rely on it, even if in some cases the correlation coefficient is big. For instance, at the level of region, I observed that there is a big correlation between the production of cereals and the number of employees who are working in the domain of bicycle's construction. Even if this may

be possible in real life, the conclusion was that I could introduce a large bias if this driver is in the model. Logically the number of employees working in field of bicycle's construction should not explain the production of cereals very well. I also implemented forward stepwise regression in order to help me to distinguish between the drivers. Unfortunately, the selected drivers are overfitting the regional model.

Following these observations, as a first step of drivers selection process, I manually choose 10 possible drivers for 22 observations, which in my opinion can describe the materials of interest (please see Appendix A).

5.4.1.1 Conditioned R^2 Criterion

The assumption done is that data does not present any outlier. In the case of territorial breakdown of French regions, an outlier represents a region, which does not have the same characteristics as the majority, such that it can not be described by the same type of model as the rest of the regions.

The coefficient of determination denoted R^2 gives some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1 indicates that the regression line perfectly fits the data. The general formula for R^2 is:

$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$, where
$SS_{res} = \sum_i (Y_i - Y_i^*)^2$ – the regression sum of squares;
$SS_{tot} = \sum_i (Y_i - \bar{Y_i})^2$ – the total sum of squares;
Y_i – the data set values with $i = 1 \cdots n$;
$\bar{Y_i} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ – the mean of observed data;
Y_i^* – estimated value for Y_i.

The estimators are calculated by ordinary least-squares regression: that is, by minimizing the difference between estimated and true outcomes $Y_{r_i}^* - Y_{r_i}$, such that the condition $\sum_j \lambda_j = 1$ is fulfilled. In this case R^2 increases as I increase the number of variables in the model (R^2 is monotone increasing with the number of variables included, i.e. it will never decrease). This illustrates a drawback to one possible use of R^2, where one might keep adding variables to increase the R^2 value.

The goal is to find the best model based on R^2 criterion, but in the same time I do not want to fall in the trap of choosing the most complex model, just because R^2 increases with every added driver. Therefore the following procedure is proposed:

1. compute all the models with every possible X^j combination:
   $(\{X^1\}, ..., \{X^j\}; \{X^1, X^2\}, \{X^1, X^3\}, ..., \{X^{j-1}, X^j\}; ...; \{X^1, X^2, ..., X^j\})$, such that all the models with $1, 2, \cdots, m$ subset of drivers are obtained . For each model among all the possible combinations, calculate R^2 and keep the model which has the highest R^2.

2. for each best model found at step 1) with $1, 2, \cdots, m$ subset of drivers, do:

   - for a certain number of times do:
     - randomly generate a driver, X^{random}
     - change iteratively each X^j with X^{random} by keeping the others fixed
   - compute the mean, $\bar{R^2}$ of each model founded in this way
   - from all the models which have X^{random} instead of a chosen driver within those $1, 2, \cdots, m$ subset of selected drivers, keep the model which has the highest $\bar{R^2}$

3. starting with the maximum number of drivers, compare each model from step 1) with each one found at step 2)

4. keep the model which verifies the condition $R^2 \geq \bar{R^2} + \alpha$

For example, supposing that the maximum number of drivers is 2; A, B being the drivers and *Random* is the additional driver with its values randomly generated. One is looking to find the best model that can be described with the drivers A and B such that the R^2's drawback is covered. The chosen model will be $M(A, B)$ if the following condition is verified:

$$R^2_{M(A,B)} > \max(R^2_{M(A,Random)}, R^2_{M(B,Random)}) + \alpha$$

Otherwise, the algorithm will find the best model with only one driver within the possible chosen ones.

Alpha, α should be setted such that to reflect that the selected model is better that the one formed with randomly generated drivers. In other words, the model should not be selected by chance. The example in Figure 5.3 clarifies how α should be selected.

5.4.1.2 AIC - based criterion & RANSAC procedure

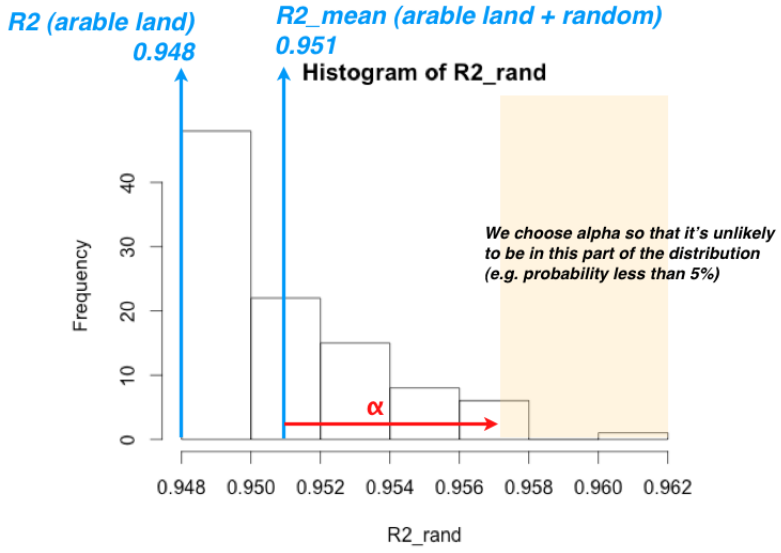
The assumption done is that up to $y\%$ of the data are outliers.

Supposing there are n observations of the response variable Y and of m drivers X^j: $Y_i, i = 1 \cdots n$ and $X^j_i, i = 1 \cdots n, j = 1 \cdots m$.

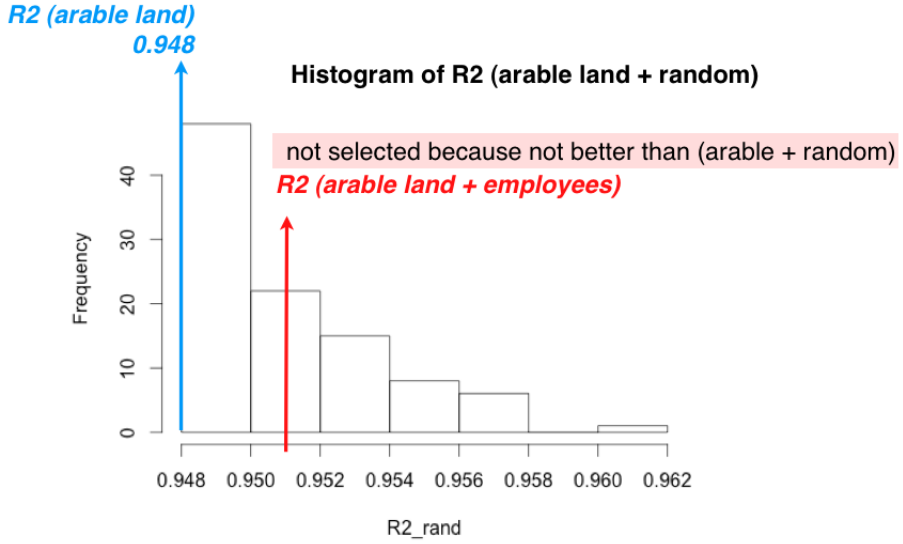
The idea is to explain the response variable by a linear model in an appropriate subset of m_s drivers amongst the given m ones. Let $\{j_k, k = 1 \cdots m_s\}$ be the set of indices of the selected drivers: $j_k \in \{1, \cdots, m\}, k = 1 \cdots m_s$.

Then, the linear model is:

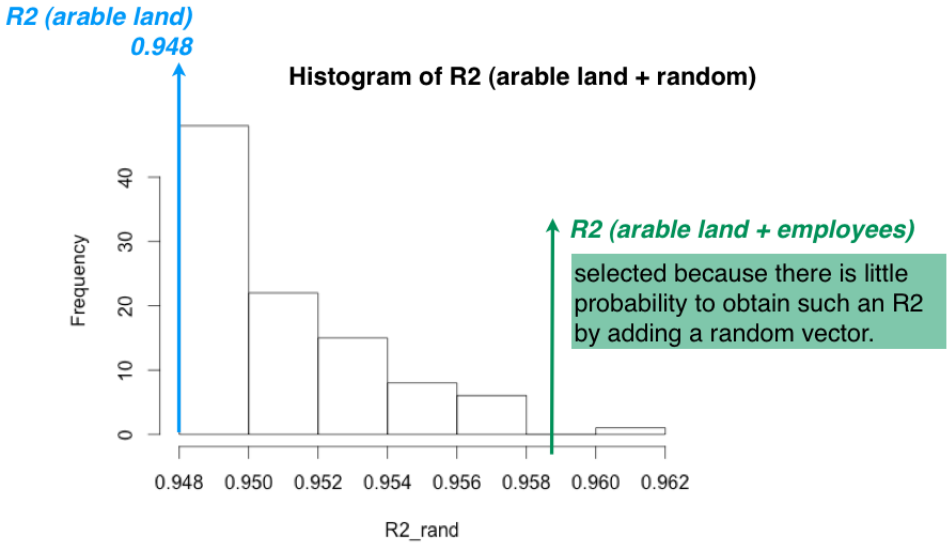
$$Y_i = \sum_{k=1}^{m_s} \lambda_{j_k} X_i^{j_k}$$



(a)



(b)



(c)

Figure 5.3: Example of alpha selection

Further, the following constraint of the coefficients must be fulfilled:

$$\sum_j \lambda_j = 1 \tag{10}$$

In principle, estimating model coefficients for a specific set of drivers can be done by a constrained least squares formulation, where the objective function is:

$$\sum_i \left(Y_i - \sum_{k=1}^{m_s} \lambda_{j_k} X_{i,j_k} \right)^2$$

and the constraint $\sum_j \lambda_j = 1$ must be respected (cf. sub-section 5.4).

The potential problem, as mentioned above, is outliers in the data, in which case the model coefficients are likely to be badly estimated.

RANSAC procedure

A solution to this problem is the RANSAC (RANdom SAmple Consensus) [21] procedure. To solve the constrained least squares problem, it is needed at least $m_s - 1$ observations. The following procedure is repeated for S times, where S ideally should be calculated depending on m_s.

1) Randomly select $m_s - 1$ observations i and solve the constrained least squares problem. This gives a hypothesis for the model coefficients λ.

2) Evaluate the hypothesis: compute the residuals for all observations (for the $m_s - 1$ observations used to compute the model here, the residuals should be zero). Compute for instance, the median of the residuals' absolute values. If the median is lower than the current best one for the considered set of drivers, keep the new computed model and its median.

3) An alternative to the median as quality measure, and which is more adapted for usage in AIC-type [8] model selection (see below) is the sum of squared residuals, computed over the x residuals with smallest absolute value (x being chosen such that $x \leq n - y$ assuming that there are at most y outliers in the data).

The outcome of this procedure is a hypothesis of the model coefficients for the considered set of drivers, as well as a quality measure for this set of drivers.

AIC criterion

For every set of drivers one wants to consider for the problem, the procedure described above can be used to compute the model coefficients for these drivers as well as a quality measure of the set of drivers.

The two possible quality measures mentioned (median of absolute residuals or sum of squares of the x residuals with smallest absolute value) can be used to select either:

- for one set of drivers, select among different estimates for the model coefficients (this is what is done inside RANSAC [21]).
- for different sets of drivers (and model coefficients for each set), select the best set.

The latter only makes sense if the considered sets of drivers have the same cardinality. Also, another problem is that R^2 increases as the number of variables in the model increases (cf. 5.4.1.1).

Therefore a model selection criterion that takes into account the goodness of fit (above quality measures) and the number of drivers (the "size" of a model) must be used. An AIC-like criterion [8] for a set of m_s drivers would be:

$$2m_s + x \ln(\bar{R})$$

where $\bar{R}$ is the sum of squares of the x residuals with smallest absolute values.

General Procedure

A general procedure would take as input all sets of drivers one would like to test. For each set, model coefficients are estimated using RANSAC. The best set is determined as in 5.4.1.2.

A problem can be the complexity (too many sets of drivers to test). A more efficient approximate procedure is as follows. One sequentially adds one driver after the other, each time taking the best driver among the ones not used yet.

1. Initial number of drivers is $m_s = 0$.
2. For $m_s = 1$ to m do
   - For every driver X^j not yet used, do
     - Consider the set composed of the $m_s - 1$ already selected drivers, and driver X^j
     - Estimate model coefficients and a quality measure (cf. 5.4.1.2).
   - Select the best new driver, as in 5.4.1.2. Let its index be $\hat{X}^j$.
   - Add the driver to the set of selected drivers: set $j_{m_s} = \hat{X}^j$.

The implementation of the algorithms can be found in Appendix B.

6 Tests and results: Applying the downscaling hypothesis

In this section, I present and discuss the results of territorial breakdown of French regions aiming to compare the estimated data with the real one for department level.

I want to test the downscaling hypothesis: the explanatory model of the response variable remains the same when the scale changes.

Applying the downscaling hypothesis for French regions and departments implies that:

- the model at regional level is also valid for department level;
- if the department data is available, the validity of the hypothesis can be tested.

By verifying the downscaling hypothesis at department level, I am considering that if it holds for department scale then it is true also for the other administrative territories as cities, communes, etc. In other words, if it is true, then I can use the regional model to estimate the needed data not only for departments but also for communes or cities.

Further I show the models for selected materials in 5.4.1 at region level and I am checking if the downscaling hypothesis is kept also at department level.[1] Although the downscaling hypothesis verification at department level depends on the chosen drivers at region level, mathematically it has the following formulation, which prove that this type of approach holds:

Applying the estimation model to all D_i departments in the region r_i means:

$$\sum_{k=1}^{D_i} Y^*_{d_{i,k}} = \sum_{k=1}^{D_i}\sum_{j=1}^{m} \lambda_j X^j_{d_{i,k}} = \sum_{j=1}^{m} \lambda_j \underbrace{\sum_{k=1}^{D_i} X^j_{d_{i,k}}}_{\text{this is equal to } X^j_{r_i}} = \sum_{j=1}^{m} \lambda_j X^j_{r_i} = Y^*_{r_i}$$

Hence:

$$\sum_{k=1}^{D_i} Y^*_{d_{i,k}} = Y^*_{r_i}$$

In the case of second algorithm explained in 5.4.1.2, the testing procedures are done in the context that up to 20% of the data are outliers.

Best model at region level using Conditioned R^2 Approach		
Drivers	Coefficients	R^2
Non-irrigated arable land	0.863	0.968
Number of employees working in the industry of cereals	0.138	
Best model at region level using AIC-based criterion & RANSAC Approach		
Drivers	Coefficients	R^2
Non-irrigated arable land	0.747	0.995
Number of employees working in the industry of cereals	0.260	
Farmers population	−0.189	
Number of employees working in the milling industry	0.351	
Population	−0.160	
Number of employees working in the industry of starches' manufacture	−0.01	

Table 6.1: Best Model for Cereals

R^2 obtained by applying the regional model to department data	
Conditioned R^2 Approach	0.928
AIC-based criterion & RANSAC Approach (the outliers from department data are removed)	0.956
AIC-based criterion & RANSAC Approach (the departments from outliers region are removed)	0.892

Table 6.2: Verification of Downscaling Hypothesis for Cereals

6.1 Cereals

The proposed algorithms select "Non-irrigated arable land" as being the most important driver for estimating the cereals production at regional scale, while the other drivers are less important. Based on the coefficient's value of this driver, I can conclude that it has a big importance in explaining the production of cereals. The R^2 coefficient shows how good the model fits the data and also can be interpreted as the percentage of the variability of the data explained by the model. In both cases the R^2 value of model found for cereal's production at regional is above 95% (Table 6.1).

In order to check if the downscaling hypothesis holds, I observe how good the model found at regional level explains the data at department level. In the case of second algorithm, I perform two types of tests: for the first one, I remove from the department dataset those departments that are consider outliers; while for the second test I remove those departments, which belong to the regions that are consider outliers for the region dataset. These tests show that the performance of the regional model for the downscaling verification procedure is above 89% (Table 6.2). Therefore, I can emphasis that for cereals the proposed methodology works.

Best model at region level using Conditioned R^2 Approach		
Drivers	Coefficients	R^2
Number of employees working in the domain of fruit trees cultivation	0.832	0.934
Number of employees working in the process of fruits transformation and conservation	0.168	
Best model at region level using AIC-based criterion & RANSAC Approach		
Drivers	Coefficients	R^2
Number of employees working in the domain of fruit trees cultivation	0.697	0.985
Number of employees working in the process of fruits transformation and conservation	0.227	
Number of employees working in the area of en - gross fruit commerce	0.121	
Pastures	-0.047	

Table 6.3: Best Model for Fruits

R^2 obtained by applying the regional model to department data	
Conditioned R^2 Approach	0.780
AIC-based criterion & RANSAC Approach (the outliers from department data are removed)	0.846
AIC-based criterion & RANSAC Approach (the departments from outliers region are removed)	0.725

Table 6.4: Verification of Downscaling Hypothesis for Fruits

6.2 Fruits

In the case of fruits production at regional level, the most important driver is "Number of employees working in the domain of fruit trees cultivation". The R^2 value of the model is above 90% (Table 6.3), which underlines that the model fits very well the data even if exist or no outliers.

Based on the values of R^2 (Table 6.4) and if I consider that a good estimation of department data using the region model is above the threshold of 70%, then I can infer that the downscaling assumption holds in the case of fruits. For the first algorithm, the model found at region level explains 78% of the data. In the case of the second algorithm, after I removed the outliers from the department data, 84% of the real data is explained by the regional model. Further in the case of the second algorithm, but removing the departments from the regions which are consider outliers, the model performance drops at 72%. This difference prove that at department level, there are some points which are not consider outliers at region level.

Even if the downscaling verification result of the Conditioned R^2 algorithm is better than the latest one of the second algorithm, this does not prove that the problem of outliers should not be taking in consideration, especially because the outliers depend on model selection.

6.3 Grapes

The same analysis has been done also for grapes and the results can be seen in the Table 6.5 and Table 6.6:

[1]The values of the coefficients have been truncated, but the original ones respect the imposed condition: $\sum_i \lambda_i = 1$

Best model at region level using Conditioned R^2 Approach		
Drivers	Coefficients	R^2
Vineyards	1	0.890
Best model at region level using AIC-based criterion & RANSAC Approach		
Drivers	Coefficients	R^2
Vineyards	0.772	0.999
Number of employees working in the industry of champagne making	0.181	
Number of employees working in the field of viticulture industry	0.046	
Number of employees working in the field of liquor manufacturing industry	−0.0006	

Table 6.5: Best Model for Grapes

R^2 obtained by applying the regional model to department data	
Conditioned R^2 Approach	0.889
AIC-based criterion & RANSAC Approach (the outliers from department data are removed)	0.994
AIC-based criterion & RANSAC Approach (the departments from outliers region are removed)	0.976

Table 6.6: Verification of Downscaling Hypothesis for Grapes

"Vineyards" is the most important driver even if I apply or not outliers elimination algorithm. In the case that outliers are presented in the model, the R^2 value is 89%, while after the outliers are removed, the model performance grows up to 99% (Table 6.5).

The region model applied to the department data after the outliers are removed or after the departments from outliers region are eliminated has a performance above 97% (Table 6.6), which is much higher than 88%, founded in the case of first algorithm model validation. These results prove the validity of downscaling hypothesis in the case of grapes production dataset, if I choose a validation threshold above 88%.

6.4 Potatoes

In the case of potatoes production estimation, the model found at regional level has a good performance of 85.9% for the first case and 97.4% in the case that outliers are eliminated (Table 6.7).

It is important to observe that both algorithms select the same model, but the values of the coefficients are not the same. This difference is produced because, for the second case, since the algorithm is not enough robust to the outliers and because in the dataset can be more than 20% outliers, an additional AIC-based criterion & RANSAC procedure is applied at the level of region, using the selected drivers.

In Table 6.8 are presented the results of the downscaling verification tests. It can be noticed that the problem of outliers is not the only challenge in this analysis. By eliminating the outliers from the department data and applying the region model, the downscaling approach seems to hold and the region model manages to properly estimate more than 82% of the real department data. However, by first

Best model at region level using Conditioned R^2 Approach		
Drivers	Coefficients	R^2
Non-irrigated arable land	0.509	0.859
Number of employees working in the field of potatoes processing and preserving	0.776	
Number of employees working in additional services related to farming	−0.385	
Best model at region level using AIC-based criterion & RANSAC Approach		
Drivers	Coefficients	R^2
Non-irrigated arable land	0.441	0.974
Number of employees working in the field of potatoes processing and preserving	0.718	
Number of employees working in additional services related to farming	−0.160	

Table 6.7: Best Model for Potatoes

R^2 obtained by applying the regional model to department data	
Conditioned R^2 Approach	0.113
AIC-based criterion & RANSAC Approach (the outliers from department data are removed)	0.823
AIC-based criterion & RANSAC Approach (the departments from outliers region are removed)	−0.051

Table 6.8: Verification of Downscaling Hypothesis for Potatoes

eliminating the departments from the regions which are consider outliers and after applying the model to department data, the test of the downscaling verification show a negative R^2 equal to -0.051. As a consequence of this result, the linear model found at region level does not hold also at department level. Therefore a non-linear model should be taking in consideration for potatoes dataset.

6.5 Vegetables

Best model at region level using Conditioned R^2 Approach		
Drivers	Coefficients	R^2
Number of employees working in the process of vegetables transformation and conservation	0.523	0.857
Number of employees working in the area of en - gross fruits and vegetables commerce	0.815	
Number of employees working in the field of potatoes processing and preserving	0.776	
Best model at region level using AIC-based criterion & RANSAC Approach		
Drivers	Coefficients	R^2
Number of employees working in agricultures	0.687	0.980
Number of employees working in the process of vegetables transformation and conservation	0.312	

Table 6.9: Best Model for Vegetables

Based on the results from Table 6.9, I can infer that the model at regional level for vegetables production performs very good and it has an accuracy above 98%, if the outliers from regional data are removed.

For the downscaling verification process, if the outliers from department data are eliminated then

R^2 obtained by applying the regional model to department data	
Conditioned R^2 Approach	0.396
AIC-based criterion & RANSAC Approach (the outliers from department data are removed)	0.818
AIC-based criterion & RANSAC Approach (the departments from outliers region are removed)	0.640

Table 6.10: Verification of Downscaling Hypothesis for Vegetables

the model performance is above 80%. Although if the departments belonging to the regions, which are considers outliers are dropped, then the model performance applied to department data decreases until 64% (Table 6.10). Therefore, I can deduce that the downscaling hypothesis is verified for 64% of the real department data, while for the rest of 36% of the cases, the department data has a different particularity and it can not be characterized using a linear model.

Nevertheless, by eliminating the outliers, the model estimates better than in the situation that the outliers are present in the dataset, where the R^2 value is 0.396.

6.6 Total Wood

Best model at region level using Conditioned R^2 Approach		
Drivers	Coefficients	R^2
Coniferous forests	1.055	0.961
Mixed forests	−0.9	
Number of employees working in forest exploitation	0.240	
Number of employees working in in the domain of sawmilling and planing of wood	0.604	
Best model at region level using AIC-based criterion & RANSAC Approach		
Drivers	Coefficients	R^2
Number of employees working in the domain of sawmilling and planing of wood	0.535	0.964
Number of employees working in forest exploitation	0.619	
Number of employees working in wooden containers manufacture	−0.12	
Hardwood forests	0.103	
Number of employees working in wood panels manufacture	−0.08	
Forest and bush vegetation changing	−0.048	

Table 6.11: Best Model for Total Wood

In the case of total production of wood, both algorithms perform very well, by selecting a model which describes more than 96% of the real wood production at the level of region (Table 6.11).

By applying the regional model to the department data, more than 70% of the departemntal wood production is well explained (Table 6.12).

In the case that outliers are eliminated from the department dataset, by appling the second algorithm, the model performance grows up to 87.4%, which is much better that one of the first model given by the *Conditioned* R^2 algorithm. Although if the departments from outliers regions are removed from the department dataset before computing R^2, then the model is capable to explain only 74.1% of the

real dataset. This difference is produced because, in the department dataset exists points that are not included in the regions, which are consider oultiers.

Therefore, if I set the model performance threshold at 70%, I can infer that the donwscaling appraoch is veryfied also at department level.

R^2 obtained by applying the regional model to department data	
Conditioned R^2 Approach	0.705
AIC-based criterion & RANSAC Approach (the outliers from department data are removed)	0.874
AIC-based criterion & RANSAC Approach (the departments from outliers region are removed)	0.741

Table 6.12: Verification of Downscaling Hypothesis for Total Wood

6.7 Outliers Analysis

This section analyzes the outliers from both region and department data. Based on the downscaling hypothesis and taking in consideration the *AIC-based criterion & RANSAC* approach I perform this analysis for the worst result, which is in the case of *Potatoes* dataset and for *Vegetables* dataset, which has better results than *Potatoes* dataset. For this two cases, I will present which are the outliers from department dataset and which are the departments from outliers regions.

For potatoes dataset, the departments which are consider to be outliers or that are belonging to outliers regions, can be seen in Table 6.13, respectively Table 6.14.

The common departments, in the case of running two verification tests for potatoes dataset, can be observed in the Table 6.15. The big difference, between R^2s (0.823, -0.051) of the two tests performed using the model given by *AIC-based criterion & RANSAC* approach, is due to the fact that there is a substantial number of points in department dataset which cannot be explained by a linear model. Much more, only 6 departments out of 17 belong to the outliers regions and to outliers departments. In this case it is difficult to prove that the downscaling hypothesis holds at department level, giving that the R^2 of the model applied to the department data set is significant small (-0.051).

In the case of vegetables dataset, the outliers from department data can be seen in Table 6.16, while the departments from outliers regions can be observed in Table 6.17.

By performing two verification test for downscaling hypothesis at the department level using *AIC-based criterion & RANSAC* approach, the common departments, which are classified as being outliers can be observed in Table 6.18.

The difference between the R^2s (0.818, 0.640) in the case of verifying the downscaling hypothesis is of 0.178, which prove that the outliers regions include departments, which are not in the outliers regions, but they are important for the model performance.

Although, in the worst case the model find at region level for vegetables dataset after the departments from outliers regions are removed can correctly estimate 64% of the data. In this sense, I can deduce that the downscaling hypothesis verification does not have a very satisfactory result, but for 64% of the

Potatoes Dataset			
Outliers from department dataset for potatoes			
Dataset Index	Code Department	Department's Name	Region's Name
5	FR106	SEINE-SAINT-DENIS	Ile-de-France
9	FR212	AUBE	Champagne-Ardenne
11	FR214	HAUTE-MARNE	
12	FR221	AISNE	Picardie
13	FR222	OISE	
14	FR223	SOMME	
16	FR232	SEINE-MARITIME	Haute-Normandie
17	FR241	CHER	Centre(FR)
18	FR242	EURE-ET-LOIR	
29	FR264	YONNE	Bourgogne
30	FR301	NORD	Nord-Pas-de-Calais
31	FR302	PAS-DE-CALAIS	
35	FR414	VOSGES	Lorraine
46	FR515	VENDEE	Pays de la Loire
47	FR521	COTES-DARMOR	Bretagne
48	FR522	FINISTERE	
50	FR524	MORBIHAN	
54	FR534	VIENNE	Poitou-Charentes
63	FR624	GERS	Midi-Pyrenees

Table 6.13: Outliers from department dataset in the case of potatoes

cases it works well.

I can conclude that for some of the presented materials, by fitting a linear model at regional level, overcoming the problem of outliers and testing if the downscaling hypothesis holds, it represents a good method for material's estimation. However, in some cases the linear model is not a good approach to estimate the data at a sub-national level.

The downscaling methodology can be successfully used on other type of materials with the condition that each material depends on different set of drivers, which must be carefully chosen. The obtained models for the studied materials can be applied with success on other territories, which have a similar administrative structure to the French one.

Potatoes Dataset			
Outliers from department dataset, which are included in outliers regions			
Dataset Index	Code Department	Department's Name	Region's Name
12	FR221	AISNE	Picardie
13	FR222	OISE	
14	FR223	SOMME	
15	FR231	EURE	Haute-Normandie
16	FR232	SEINE-MARITIME	
32	FR411	MEURTHE-ET-MOSELLE	Lorraine
33	FR412	MEUSE	
34	FR413	MOSELLE	
35	FR414	VOSGES	
60	FR621	ARIEGE	Midi-Pyrenees
61	FR622	AVEYRON	
62	FR623	HAUTE-GARONNE	
63	FR624	GERS	
64	FR625	LOT	
65	FR626	HAUTES-PYRENEES	
66	FR627	TARN	
67	FR628	TARN-ET-GARONNE	

Table 6.14: Outliers from department dataset, which are included in outliers regions in the case of potatoes

Potatoes Dataset		
Common outliers from department dataset for potatoes		
Dataset Index	Department's Name	Region's Name
12	ISNE	Picardie
13	OISE	
14	SOMME	
16	SEINE-MARITIME	Haute-Normandie
35	VOSGES	Lorraine
63	GERS	Midi-Pyrenees

Table 6.15: Common outliers departments in the case of potatoes

7 Conclusion

Material Flow Analysis (MFA) is a powerful tool, which can help the analysts to understand the metabolism of a territory.

Even if a large number of MFA studies have been done at the national level, studies focusing at the regional or local level are very limited, and a standardized method does not yet exist. The available studies show the importance of material flows to regional and urban metabolism, even if it is very difficult to obtain the proper data at these levels.

Therefore the main constraint on the attempt to calculate the material flow of a small territory (e.g.:

Vegetables Dataset			
Outliers from department dataset for vegetables			
Dataset Index	Code Department	Department's Name	Region's Name
12	FR221	AISNE	Picardie
14	FR223	SOMME	
18	FR242	EURE-ET-LOIR	Centre (FR)
20	FR244	INDRE-ET-LOIRE	
24	FR252	MANCHE	Basse-Normandie
30	FR301	NORD	Nord-Pas-de-Calais
31	FR302	PAS-DE-CALAIS	
42	FR511	LOIRE-ATLANTIQUE	Pays de la Loire
47	FR521	COTES-D'ARMOR	Bretagne
48	FR522	FINISTERE	
52	FR532	CHARENTE-MARITIME	Poitou-Charentes
56	FR612	GIRONDE	Aquitaine
57	FR613	LANDES	
58	FR614	LOT-ET-GARONNE	
59	FR615	PYRENEES-ATLANTIQUES	
76	FR716	RHONE	Rhone-Alpes
84	FR812	GARD	Languedoc-Roussillon
91	FR824	BOUCHES-DU-RHONE	Provence-Alpes-Cote-d'Azur
92	FR825	VAR	

Table 6.16: Outliers from department dataset in the case of vegetables

department, city) is given by the difficulty to quantify the amounts of products, because at this level no statistical data is available.

In this paper, the introduced methodology aims to statistically estimate the MFA data for a specific level of a territory. This type of work is integrated in the context of ecological accounting and material flow analysis, which represents one of the main interests of the STEEP team regarding the problem of long-term sustainability at sub-national scales.

The approach is looking for the best model that can fit the data explained by the other variables, called drivers, which are available at different scales. In order to build the model I must properly select the drivers. Correlation studies prove that in some cases what is observed as being a good mathematical relationship it is not also true in the real world (e.g.: correlation between production of cereals and employees working in bicycle industry).

The downscaling technique is applied for French regions. The assumption that I want to test is if the model found at regional level is also true at department level. The tests prove that the hypothesis holds, and obtained models describe the data with a very high accuracy. Model and drivers selection relies on two implemented methodologies. The first one makes a compromise between the complexity of the model and its performance, without taking in consideration the presence of outliers. The last one, takes in consideration the outliers and it also makes a compromise between the goodness of fit and the number of drivers of a model. During the test, it has been proved that outliers can misled the analysis.

Vegetables Dataset			
Outliers from department dataset, which are included in outliers regions			
Dataset Index	Code Department	Department's Name	Region's Name
8	FR211	ARDENNES	Champagne-Ardenne
9	FR212	AUBE	
10	FR213	MARNE	
11	FR214	HAUTE-MARNE	
42	FR511	LOIRE-ATLANTIQUE	Pays de la Loire
43	FR512	MAINE-ET-LOIRE	
44	FR513	MAYENNE	
45	FR514	SARTHE	
46	FR515	VENDEE	
55	FR611	DORDOGNE	Aquitaine
56	FR612	GIRONDE	
57	FR613	LANDES	
58	FR614	LOT-ET-GARONNE	
59	FR615	PYRENEES-ATLANTIQUES	
83	FR811	AUDE	Languedoc-Roussillon
84	FR812	GARD	
85	FR813	HERAULT	
86	FR814	LOZERE	
87	FR815	PYRENEES-ORIENTALES	

Table 6.17: Outliers from department dataset, which are included in outliers regions in the case of vegetables

Vegetables Dataset		
Common outliers from department dataset for vegetables		
Dataset Index	Department's Name	Region's Name
42	LOIRE-ATLANTIQUE	Pays de la Loire
56	GIRONDE	Aquitaine
57	LANDES	
58	LOT-ET-GARONNE	
59	PYRENEES-ATLANTIQUESGARD	
84	LOIRE-ATLANTIQUE	Languedoc-Roussillon

Table 6.18: Common outliers departments in the case of vegetables

Further this application can be adjusted for different materials and territories.

8 Further Work

Regarding the problem of drivers selection, I would like to automatically choose the proper drivers for data analysis, by implementing an algorithm that can detect from the entire range of possible drivers, only those which can describe the data.

For the model selection, I am interested in implementing a better algorithm, which can find the best model in a robust cross-validation manner. [32]

A linear model, may not be a good technique for estimating the different types of materials, such that further it would be interesting to adapt the proposed methodology to non-linear data modeling.

Bayesian network represents another approach in determining how the materials and drivers interact and it can give an idea of which drivers could cause/explain the production of a specific material in a given territory.

References

[1] AGRESTE, http://agreste.agriculture.gouv.fr/.

[2] AgriMer, http://www.franceagrimer.fr/.

[3] CLC (Corine Land Cover for France) http://sd1878-2.sivit.org/.

[4] http://steep.inrialpes.fr/.

[5] http://www.ercim.eu/.

[6] http://www.inria.fr/.

[7] INSEE, http://www.insee.fr/fr/.

[8] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723, 1974.

[9] X. Bai. Industrial ecology and the global impacts of cities. *Journal of Industrial Ecology*, 11(2), 2007.

[10] S. Barles. Urban Metabolism of Paris and its Region. *Journal of Industrial Ecology*, 13(6):898–913, 2009.

[11] BFF (Best Foot Forward). City limits: A resource flow and ecological footprint analysis of Greater London, 2002.

[12] C. Binder. From material flow analysis to material flow management Part I: Social sciences modeling approaches coupled to MFA. *Journal of Cleaner Production*, 15:1596–1604, 2007.

[13] S. Boyden, S. Millar, K. Newcombe, and B. O'neill. The ecology of a city and its people: The case of Hong Kong. *Australian National University Press*, 1981.

[14] P. Brunner, H. Daxbeck, and P. Baccini. Industrial metabolism at the regional and local level: A case-study on a Swiss region. In R. U. Ayres and U. E. Simonis, editors, *Industrial metabolism: Restructuring for sustainable development*, Tokyo: United Nations University Press, 1994.

[15] F. Burström, N. Brandt, B. Frostell, and U. Mohlander. Material Flow Accounting and Information for Environmental Policies in the City of Stockholm. In S. Bringezu, M. Fischer-Kowalski, R. Kleijn, and V. Palm, editors, *Analysis for action: Support for policy towards sustainability by material flow accounting*, pages 136–145, Wuppertal, Germany: Wuppertal Institute for Climate, Environment and Energy, 1998.

[16] Jean-Yves Courtonne. Étude des flux de matières et d'énergie: région Rhône-Alpes, département de l'Isère, bassin d'emploi de Grenoble. [Material and energy flow study: Rhône-Alpes region, Isère department, Grenoble and its industrial area]. Master's thesis, STEEP Team - INRIA, 2011.

[17] H. E. Daly. Beyond growth: The economics of sustainable development. *Boston Journal*, 1996.

[18] H. Daxbeck, C. Lampert, L. Morf, R. Obernosterer, H. Rechberger, I. Reiner, and P.H. Brunner. The anthropogenic metabolism of the city of Vienna. In S. Bringezu, M. Fischer-Kowalski, R. Kleijn, and V. Palm, editors, *Regional and national material flow accounting: from paradigm to practice*, pages 249–254, Wuppertal, Germany: Wuppertal Institute for Climate, Environment and Energy, 1997.

[19] Suren Erkman and GEDEC. Ecologie industrielle à Genéve. [Industrial ecology of Geneva], 2005.

[20] M. Faist Emmenegger and R. Frischknecht. Métabolisme du canton de Genève. Phase 1. [Metabolism of Geneva. Phase 1]. *Uster: ESU Service*, 2003.

[21] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Graphics and Image Processing*, 24(6):381–395, 1981.

[22] T. E. Graedel. Industrial ecology and the ecocity. *The Bridge*, 29(4):4–9, 1999.

[23] M. Hammer, S. Giljum, F. Luks, and M. Winkler. Die ökologische Nachhaltigkeit regionaler Metabolismen: Materialflussanalysen der Regionen Hamburg, Wien und Leipzig. [Ecological sustainability or regional metabolisms: Material flow analyses of the regions of Hamburg, Vienna and Leipzig]. *Natur und Kultur*, 7(2):62–78, 2006.

[24] C. Hendriks, D. Müller, S. Kytzia, P. Baccini, and P. Brunner. Material flow analysis: A tool to support environmental policy decision making. Case studies on the city of Vienna and the Swiss lowlands. *Local Environment*, 5(3):311–328, 2000.

[25] J. Kovanda, J .and Weinzettel and T. Hak. Analysis of regional material flows: The case of the Czech Republic. *Journal of Resources, Conservation and Recycling*, 53(3), 2009.

[26] S. Niza and P. Ferrao. Material flow accounting tools and its contribution for policy making.Paper presented at the International Conference of the European Society for Ecological Economics. 2005.

[27] S. Niza, L. Rosado, and P. Ferrao. Methodological Advances in Urban Material Flow Accounting Based on the Lisbon Case Study. *Journal of Industrial Ecology*, 13(3), 2009.

[28] R. Obernosterer. Urban metal stocks: Future problem or future resource? Substance flow and stock analysis as a tool to achieve sustainable development. Paper presented at the International Conference Regional Cycles: Regional Economy Towards Sustainability. 2002.

[29] E. P. Odum. Ecology. *Orlando, Florida, USA: Holt, Rinehart, and Winston*, 1975.

[30] Statistical Office of the European Union. EUROSTAT. Economy-wide material flow accounts and derived indicators: A methodological guide, 2001.

[31] P. Riehmann, M. Hanfler, and B. Froehlich. Interactive Sankey Diagrams. IEEE Symposium on Information Visualization. 2005.

[32] E. Ronchetti, C. Field, and W. Blanchard. Robust Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*, 92(439), 1997.

[33] A. Wolman. The Metabolism of Cities. *Scientific American*, 213(3):179–188, 190, 1965.

ANNEXES

A First appendix

Prooving that $\sum_{j=1}^{m} \lambda_j = 1$:

Known:

$$Y_{r_i} = \sum_{j=1}^{m} \lambda_j X_{r_i}^j \tag{1}$$

After normalization:

$$\sum_{i=1}^{n} Y_{r_i} = 1 \tag{2}$$

$$\sum_{i=1}^{n} X_{r_i}^j = 1 \tag{3}$$

By summing up all the Y_{r_i} the first equation becomes:

$$\sum_{i=1}^{n} Y_{r_i} = \sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_j X_{r_i}^j \tag{4}$$

Knowing that $\sum_{i=1}^{n} Y_{r_i} = 1$ then:

$$1 = \sum_{i=1}^{n} Y_{r_i} = \sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_j X_{r_i}^j \tag{5}$$

$$1 = \sum_{j=1}^{m} \lambda_j \sum_{i=1}^{n} X_{r_i}^j \tag{6}$$

Because $\sum_{i=1}^{n} X_{r_i}^j = 1$, then equation (6) becomes:

$$1 = \sum_{j=1}^{m} \lambda_j * 1 \tag{7}$$

Hence:

$$1 = \sum_{j=1}^{m} \lambda_j \tag{8}$$

Code	Description	Code	Description
111	Continuous urban fabric	311	Broad-leaved forest
112	Discontinuous urban fabric	312	Coniferous forest
121	Industrial or commercial units	313	Mixed forest
122	Road and rail networks and associated land	321	Natural grasslands
123	Port areas	322	Moors and heathland
124	Airports	323	Sclerophyllous vegetation
131	Mineral extraction sites	324	Transitional woodland–shrub
132	Dump sites	331	Beaches, dunes, sands
133	Construction sites	332	Bare rocks
141	Green urban areas	333	Sparsely vegetated areas
142	Sport and leisure facilities	334	Burnt areas
211	Non-irrigated arable land	335	Glaciers and perpetual snow
212	Permanently irrigated land	411	Inland marshes
213	Rice fields	412	Peat bogs
221	Vineyards	421	Salt marshes
222	Fruit trees and berry plantations	422	Salines
223	Olive groves	423	Intertidal flats
231	Pastures	511	Water courses
241	Annual crops associated with permanent crops	512	Water bodies
242	Complex cultivation patterns	521	Coastal lagoons
243	Land principally occupied by agriculture, with significant areas of natural vegetation	522	Estuaries
244	Agro-forestry areas	523	Sea and ocean

Table A.1: CLC Surfaces

Cereals		
Response Variable	Drivers's Name	Driver's Type
Production of Cereals (in *kilotons* before normalization)	Non-irrigated arable land	Quantitative
	Population	
	Farmers population	
	Number of employees working in the industry of cereals	
	Number of employees working in the milling industry	
	Number of employees working in the industry of starches' manufacture	
	Number of employees working in the business area of canteens and restaurants	
	Number of employees working in the field of food making	
	Added value 2007	

Table A.2: Manually Selected Drivers for Cereals

Fruits		
Response Variable	Drivers's Name	Driver's Type
Production of Fruits (in *kilotons* before normalization)	Fruit trees and berry plantations	Quantitative
	Complex cultivation patterns	
	Pastures	
	Land principally occupied by agriculture, with significant areas of natural vegetation	
	Number of employees working in the domain of fruit trees cultivation	
	Number of employees working in the field of fruit and vegetables juice preparation	
	Number of employees working in the process of fruits transformation and conservation	
	Number of employees working in the area of en - gross fruit commerce	
	Number of employees working in the area of en - detailed fruit commerce	

Table A.3: Manually Selected Drivers for Fruits

Potatoes		
Response Variable	Drivers's Name	Driver's Type
Production of Potatoes (in *kilotons* before normalization)	Non-irrigated arable land	Quantitative
	Permanently irrigated land	
	Complex cultivation patterns	
	Land principally occupied by agriculture, with significant areas of natural vegetation	
	Number of employees working in crop cultivation and its associated livestock	
	Number of employees working in the domain of crop production services	
	Number of employees working in additional services related to farming	
	Number of employees working in the field of potatoes processing and preserving	

Table A.4: Manually Selected Drivers for Potatoes

Vegetables		
Response Variable	Drivers's Name	Driver's Type
Production of Vegetables (in *kilotons* before normalization)	Non-irrigated arable land	Quantitative
	Permanently irrigated land	
	Rice fields	
	Complex cultivation patterns	
	Land principally occupied by agriculture, with significant areas of natural vegetation	
	Number of employees working in agriculture	
	Number of employees working in the field of fruit and vegetables juice preparation	
	Number of employees working in the process of vegetables transformation and conservation	
	Number of employees working in the area of en - gross fruits and vegetables commerce	
	Number of employees working in the area of en - detailed fruits and vegetables commerce	

Table A.5: Manually Selected Drivers for Vegetables

Grapes		
Response Variable	Drivers's Name	Driver's Type
Production of Grapes (in *kilotons* before normalization)	Vineyards	Quantitative
	Number of employees working in the field of viticulture industry	
	Number of employees working in the field of liquor manufacturing industry	
	Number of employees working in the field of ethyl alcohol manufacturing industry	
	Number of employees working in the industry of champagne making	
	Number of employees working in the wine industry	
	Number of employees working in the production of other fermented beverages	
	Number of employees working in the brewing industry	

Table A.6: Manually Selected Drivers for Grapes

Total Wood		
Response Variable	Drivers's Name	Driver's Type
Production of Total Wood (in m^3 before normalization)	Hardwood forests	Quantitative
	Coniferous forests	
	Mixed forests	
	Forest and bush vegetation in changing	
	Number of employees working in sylviculture	
	Number of employee working in forest exploitation	
	Number of employee working in forestry services	
	Number of Employee working in the domain of sawmilling and planing of wood	
	Number of Employee working in the field of wood impregnation	
	Number of Employee working in wood panels manufacture	
	Number of Employee working in wooden containers manufacture	
	Number of Employee working in other products of wood manufacture	
	Number of Employee working in wood joinery and plastics manufacture	
	Number of Employee working in intermediate trade domain of wood and building materials	
	Number of Employee working in wholesale of wood and derived products	

Table A.7: Manually Selected Drivers for Total Wood

B Second appendix

```
normalization<-function(mat,mat_norm){
  for(j in 1:ncol(mat)){
    s=mean(mat[,j])*nrow(mat)
    for(i in 1:nrow(mat)){
      mat_norm[i,j]=mat[i,j]/s
    }
  }
  return(mat_norm)
}
```

./Normalization.R

```
# Y - response
# X - explanatory variables


Rsquared<-function(X,Y) {
  res  = optimization(X,Y)
  s<-rep(0,nrow(X))
  for(i in 1:ncol(X)){
    p<-res[i,]*X[,i]
    s<-s+p
  }
  Yhat <- s
  SSE=sum((Y-Yhat)^2)
  SST=sum((Y-mean(Y))^2)
  R2=1-SSE/SST
  return(R2)
}
```

./Rsquared.R

```
# Y - response
# X - explanatory variables

optimization <- function(X,Y) {
  l <- ncol(X)
  I <- rep(1,l)
  mydata <- list()
  mydata[[1]] <- t(X)%*%X
  mydata[[2]] <- I
  mydata[[3]] <- t(I)
  mydata[[4]] <- 0
  x<-abind(mydata[[1]],mydata[[2]])
  y<-abind(mydata[[3]],mydata[[4]])
  z<-abind(x,y,along=1)

  # adding the case when determinant of z is different than 0

  if(det(z)!= 0){
    z1<-solve(z,tol=rcond(z))
    mydata[[5]] <- t(X)%*%Y
```

```
    mydata[[6]] <- 1
    w<-abind(mydata[[5]],mydata[[6]],along=1)
    r<-z1%*%w

  }

  # adding the case when determinnat of z is not different than 0

  else{ r <- NULL}

  return(r)
}
```

./Downscaling.R

```
# Y - response
# X - explanatory variables
# Drivers - the entire set of chosen drivers


RSquared_plus_random <- function(X,Y) {
  R2_list <- NULL
  for (i in 1:100) {
      Random <- as.matrix(runif(nrow(Y),0,1))
      Random <- Random / sum(Random)
      X_plus_random <- abind(X,Random)
      R2_list <- c(R2_list,Rsquared(X_plus_random,Y))
  }
  return(mean(R2_list))
}

# Build all the possible models and keep the max of each level.
R2_max_all <- NULL
comb_max_all <- NULL
R2_plus_random_max_all <- NULL
R2_plus_random_max_all <- RSquared_plus_random(NULL,Y)
nb_drivers = ncol(Drivers)
for (level in 1:nb_drivers) {
  comb_max <- 0
  R2_max <- -100
  X_max <- NULL
  R2_plus_random_max <- -100
  combinaisons <- NULL
  combinations <- combn(1:nb_drivers,level)
  for (col_index in 1 : ncol(combinations)) {
    Xcomb <- NULL
    for (row_index in 1 : nrow(combinations)) {
      Xinter <- NULL
      Xinter <- as.matrix(Drivers[,combinations[row_index,col_index]])
      Xcomb <- abind(Xcomb,Xinter)
    }
    X <- Xcomb
    r2 = Rsquared(X,Y)
    if (r2 > R2_max) {
```

```
      R2_max = r2
      comb_max = col_index
      X_max = X
    }

  }
  R2_plus_random_max = RSquared_plus_random(X_max,Y)
  R2_max_all = abind(R2_max_all,R2_max)
  comb_max_all = abind(comb_max_all,comb_max)
  R2_plus_random_max_all = abind(R2_plus_random_max_all,R2_plus_random_max)

}

# Loop through the models and return the best one.
alpha = 0.01
for (i in nb_drivers:1) {
  if (R2_max_all[i] > R2_plus_random_max_all[i] + alpha) {
    best_model <- c(i,comb_max_all[i])
    # obtaining the indexes of the drivers
    model = combn(1:nb_drivers,best_model[[1]])[,best_model[[2]]]

    break;
  }
  else if (i == 1) {
    model <- 0
  }
}
```

./Conditioned–R2.R

```
# Y - response
# X - explanatory variables
# Drivers - the entire set of chosen drivers

Y_estimated<-function(X,lambdas){
  s<-rep(0,nrow(X))
  for(i in 1:ncol(X)){
    p<-lambdas[i,]*X[,i]
    s<-s+p
  }
  Yhat<-s
  return(Yhat)
}


# approximate number of possible outliers

remove<-floor(20/100*nrow(Y))
remove

# remain

remain<-nrow(Y)-remove
remain
```

```
25
26
27
28 # finding first best driver based on AIC - modified criterion
29
30 AIC_list = NULL
31 AIC_list_f = NULL
32
33 for(d in 1:ncol(Drivers)){
34   X = as.matrix(Drivers[,d])
35   # calculating the coefficients of the model
36   lbd = optimization(X,Y)
37   Yhat = Y_estimated(X,lbd)
38   res = abs(Y-Yhat)
39   res_sort = sort(res)
40   R = sum(res_sort[1:remain]^2)
41   m = ncol(X)
42   AIC = 2*m + remain*log(R)
43   AIC_list_f = c(AIC_list_f,AIC)
44
45 }
46
47 # determining first driver
48 index_first_dv<-which(AIC_list_f==min(AIC_list_f))
49
50
51 # computing the list of good drivers
52 X_update=as.matrix(Drivers[,index_first_dv])
53 dim = dimnames(Drivers)[[2]][index_first_dv]
54 colnames(X_update) = dim
55 Drivers_update = Drivers[,-index_first_dv]
56
57 sample=100
58
59 AIC_total = min(AIC_list_f)
60
61
62 while( ncol(Drivers_update) > 0){
63
64   nd = ncol(X_update)
65
66   point_s=nd  # depends on the number of explanatory and explained variables
67   # nd represents the number of drivers that it will be in the model
68
69
70   for(du in 1:ncol(Drivers_update)){
71     # Adding separetelly each driver from the list of the remaining ones
72     D_up = abind(X_update,Drivers_update[,du],along=2)
73     AIC_sample = NULL
74     m = ncol(X_update)
75     for(i in 1:sample){
76
77       index<-sample(nrow(Y),point_s)
78       index<-sort(index)
```

```
79
80        # computing X_s and Y_s of the sample
81
82        Y_s<-matrix(NA,nrow=length(index),ncol=1)
83        X_s<-matrix(NA,nrow=length(index),ncol=ncol(X_update)+1)
84        for(j in 1:length(index)){
85          Y_s[j,1]<-Y[index[j],]
86          for(k in 1:ncol(D_up)){
87            X_s[j,k]<-D_up[index[j],k]
88          }
89        }
90
91
92        lbd = optimization(X_s,Y_s)
93        if(length(lbd) != 0){
94          Yhat = Y_estimated(D_up,lbd)
95          res = abs(Y-Yhat)
96          res_sort = sort(res,decreasing=F)
97          R = sum(res_sort[1:remain]^2)
98          AIC = 2*m + remain*log(R)
99
100       } else { AIC = 1000}
101
102       AIC_sample = c(AIC_sample,AIC)
103
104     }
105
106     AIC_min = min(AIC_sample)
107     AIC_list = c(AIC_list,AIC_min)
108   }
109
110   # computing the index of the next best driver
111   index_best<-which(AIC_list==min(AIC_list))
112
113   if (AIC_total[length(AIC_total)] > AIC_list[index_best]){
114     AIC_total = c(AIC_total,AIC_list[index_best])
115     X_update = abind(X_update,Drivers_update[,index_best],along=2)
116     dim = c(dim,dimnames(Drivers_update)[[2]][index_best])
117     Drivers_update = as.matrix(Drivers_update[,-index_best])
118   } else{
119     Drivers_update = as.matrix(Drivers_update[,-index_best])
120   }
121   AIC_list = NULL
122
123 }
124
125
126 colnames(X_update) = dim
127
128 # finding the best drivers
129 X_selected = X_update
130
131
132
```

```
# finding the best lambdas


best_lambdas<-function(X_selected,Y){

  AIC_s = NULL

  points = ncol(X_selected)

  sample=100

  lambdas_list = list()

  for(i in 1:sample){

    index<-sample(nrow(data),points)
    index<-sort(index)

    # computing X_s and Y_s of the sample

    Y_s<-matrix(NA,nrow=length(index),ncol=1)
    X_s<-matrix(NA,nrow=length(index),ncol=ncol(X_selected))
    for(j in 1:length(index)){
      Y_s[j,1]<-Y[index[j],]
      for(k in 1:ncol(X_selected)){
        X_s[j,k]<-X_selected[index[j],k]
      }
    }


    lb = optimization(X_s,Y_s)
    lambdas_list[[i]] = lb

    if(length(lb) != 0){
      Yhat = Y_estimated(X_selected,lb)
      res = abs(Y-Yhat)
      res_sort = sort(res)
      R = sum(res_sort[1:remain]^2)
      AIC = 2*points + remain*log(R)

    } else { AIC = 1000}

    AIC_s = c(AIC_s,AIC)

    }


  index = which(AIC_s==min(AIC_s))
  lambdas_best = lambdas_list[[index]]
  return(lambdas_best)



}
```

```
187
188
189
190 # Computing the R2 taking out the outliers
191
192 X = X_selected
193 lambdas = best_lambdas(X,Y)
194 s = s<-rep(0,nrow(Y))
195
196 for (i in 1: (length(lambdas) -1) ){
197   t = lambdas[i]*X[,i]
198   s<-s+t
199 }
200   Yhat = s
201   res = abs(Y-Yhat)
202   res_o = sort(res)
203   index = order(res,decreasing = FALSE)
204   index = index[1:remain]
205   data_o = NULL
206   for (ix in 1: length(index)){
207     data_o = c(data_o,Y[index[ix],])
208   }
209   Y_o = matrix(data_o,nrow=length(index),ncol=1)
210   SST_o=sum((Y_o-mean(Y_o))^2)
211   SSE_o = sum(res_o[1:remain]^2)
212   R2_o = 1-SSE_o/SST_o
```

./Ransac–AIC.R